Fall 2023: CS 4435/5435 and DASE 4435 Data Mining

Homework 4, Due on Nov 16, 2022

NOTE:

- **(a)** This is not a programming assignment. This problem can be solved **MANUALLY**, using paper and pen or any other way (e.g. using a drawing software).
- **(b)** Since the problem below is a paper and pen exercise, make sure that your work is presented neatly, and your answers can be read clearly.
- **(c)** Asterisk (*) means it is for graduate student only and bonus for undergraduate students.

1. Compute the inverse document frequency (idf) for the terms in Table 1 for a collection of 806,791 documents. Present the answer in a tabular format as presented below. (**20 points)**

term	df_t
car	18,562
auto	8,201
insurance	13,982
hero	23,793
learn	7,396
chromosome	24,234

Table 1: The document frequency, df_t, for a set of terms is provided below

2. Compute the $\mathrm{tf} - \mathrm{idf}$ weights for the terms: car, auto, insurance, hero, learn, chromosome for each document using the idf values from *question (1)*. Present the answer in a tabular format as presented below. (**40 points**)

term	Doc1	Doc2	Doc3
car	18	5	62
auto	17	20	14
insurance	13	15	25
hero	33	57	73
learn	7	3	45
chromosome	25	23	4

Table 2: Table of term frequency, tf, values

3. As a search engine provider, a user typed the words**: auto insurance hero** into your search engine page. **Rank** the three documents in Table 2 using the ***overlap score measure*** in the order they will be presented to the user. Provide the *overlap score measure values* used to perform the document ranking in your answer. (**40 points for Undergraduates, 30 points for Graduates**)

4. What is the idf of a term that occurs once in every document? Comment on your answer. * (**10 points**)